P. LE ROY[1], *J.-M. ELSEN*[2]

[1] *INRA, Station de Génétique Quantitative, et Appliquée, 78352 Jouy-en Josas cedex*
[2] *INRA, Station d'Amélioration Génétique des Animaux, BP 27, 31326 Castanet-Tolosan cedex*

e-mail : Pascale.Leroy@dga.inra.fr

7 - Utilisation des marqueurs génétiques

Principes de l'utilisation des marqueurs génétiques pour la détection des gènes influençant les caractères quantitatifs

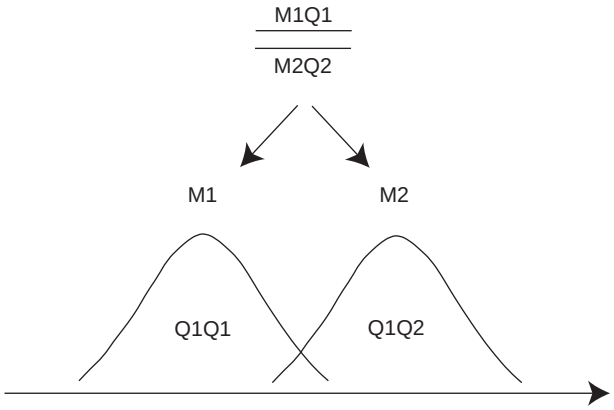
Résumé. *L'information apportée par les marqueurs génétiques est une aide précieuse pour la mise en évidence des gènes influençant les caractères quantitatifs, les QTL (pour Quantitative Trait Loci). Le principe de base est d'observer s'il y a coségrégation des allèles au marqueur et au QTL dans la descendance de reproducteurs doubles hétérozygotes. La démarche peut être étendue à la prise en compte simultanée d'informations sur plusieurs marqueurs appartenant à un même groupe de liaison : c'est la cartographie d'intervalle. Différentes méthodes statistiques sont employées pour analyser les données. Toutefois, la théorie classique des tests d'hypothèse n'étant pas complètement applicable, il est souvent fait appel aux simulations pour déterminer les seuils de rejet et les intervalles de confiance. Les protocoles mis en place sont de deux types : croisement entre populations extrêmes ou étude intra-famille de populations en ségrégation. Ils doivent être optimisés pour avoir une puissance maximale à taille de protocole fixée.*

Le modèle polygénique, utilisé classiquement en génétique quantitative, considère que les phénotypes observés résultent de l'expression d'une infinité de gènes, dont les effets individuels faibles s'additionnent, et à laquelle s'ajoute un effet du milieu. Bien que parfaitement opérationnel en sélection, ce modèle est bien sûr biologiquement faux. Il peut parfois être avantageux de le compliquer un peu en prenant explicitement en compte dans la sélection, en plus du fond polygénique habituel, l'existence d'un ou plusieurs gènes ayant un effet individuel « fort » sur le caractère (Elsen 2000, cet ouvrage). La recherche de tels gènes a été entreprise de tout temps, mais elle est récemment devenue plus efficace grâce à l'établissement de cartes génétiques pour les espèces d'élevage. Une démarche de recherche systématique a alors pu être entreprise, l'ensemble du génome étant balayé pour détecter les zones chromosomiques influençant le caractère : les Quantitative Trait Loci (QTL) (Goffinet *et al* 1994).

1 / Principe général

Pour un marqueur donné, le principe est d'observer dans la descendance d'un parent hétérozygote M1/M2 s'il existe une différence de performance moyenne selon l'allèle marqueur, M1 ou M2, transmis (figure 1). L'idée est que, si cette différence existe, elle s'explique par la ségrégation des allèles, Q1 ou Q2, en un QTL génétiquement lié au marqueur. Par exemple, dans le cas d'un croisement en retour (CR) entre deux populations P1, homozygote M1Q1/M1Q1, et P2, homozygote M2Q2/M2Q2, un père F1 (M1Q1/M2Q2) accouplé à une conjointe de P1 donne un descendant Q1/Q1 lorsqu'il transmet M1 et un descendant Q2/Q1 lorsqu'il transmet M2, en supposant l'absence de recombinaison entre M et Q.

Figure 1. *Principe général de la détection d'un QTL.*



Par rapport à ce schéma idéal, il peut exister trois limites :

- s'il y a ségrégation des allèles, au QTL et au marqueur, chez les deux parents, l'observation du mélange des phénotypes chez les descendants d'une part, et l'établissement de la transmission parent-descendant des allèles au marqueur d'autre part, sont plus difficiles ;
- si le taux de recombinaison entre le marqueur et

le QTL n'est pas nul, la différence estimée entre groupes de descendants ayant reçu M1 ou M2 diminue ;

- s'il n'y a pas déséquilibre complet d'association gamétique entre le marqueur et le QTL, c'est-à-dire s'il existe d'autres haplotypes que M1Q1 et M2Q2 dans la population, certains parents peuvent être hétérozygotes au marqueur mais homozygotes au QTL (par exemple M1Q1/M2Q1) et, par suite, leurs familles ne sont pas informatives. De plus, il peut y avoir des parents M1Q1/M2Q2 et des parents M1Q2/M2Q1, ce qui impose d'analyser les données intra-famille.

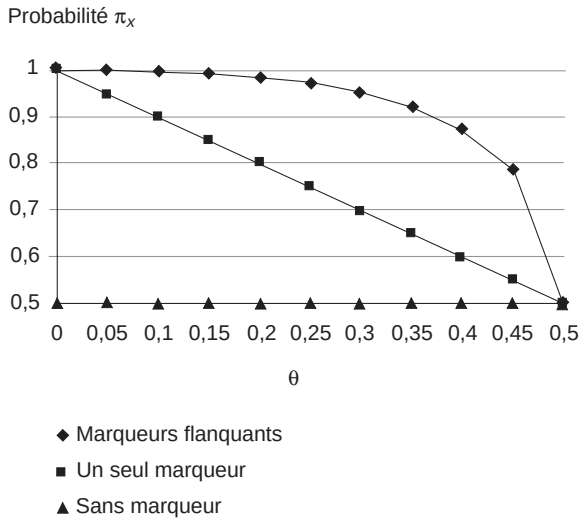
Pour cela, une analyse de variance peut être appliquée avec un modèle linéaire prenant en compte les effets du père et de l'allèle marqueur transmis intrapère (Soller et Genizi 1978). La différence estimée (Δ), entre performances moyennes des deux groupes de descendants ayant reçu l'un ou l'autre allèle de leur père, est un effet apparent du marqueur. Δ est une fonction de l'effet du QTL ($2a$ = différence entre les génotypes homozygotes Q1/Q1 et Q2/Q2) et du taux de recombinaison entre le marqueur et le QTL (θ) : $\Delta=2a(1-2\theta)$. Par cette méthode, il y a donc confusion entre la position et l'effet du QTL. L'utilisation de techniques du maximum de vraisemblance évite en théorie cet écueil, en permettant d'estimer les deux paramètres θ et a. Toutefois, il faut souligner que l'estimation précise de θ requiert une quantité importante de données.

2 / La cartographie d'intervalle

Lorsqu'une carte génétique est disponible, il est possible d'envisager une démarche de recherche systématique des QTL sur l'ensemble du génome. Les données sont alors analysées par groupe de liaison, en général par chromosome, tous les marqueurs liés pouvant informer sur la ségrégation d'un QTL dans une zone donnée. Le principe est de balayer l'ensemble du génome en testant l'hypothèse d'absence de QTL en chaque position x : c'est la cartographie d'intervalle (Lander et Botstein, 1989). L'utilisation d'informations sur deux marqueurs entourant le QTL (marqueurs flanquants) permet de discerner position et effet du QTL, les estimations des paramètres θ et a étant bien meilleures que lors de la prise en compte d'un seul marqueur à la fois. Par ailleurs, cette démarche multimarqueurs est en principe plus puissante.

En reprenant l'exemple simple du CR (P1 x P2) x P1, le parent F1 double hétérozygote est porteur de deux haplotypes marqueurs, l'un caractéristique de la population P1 (M1N1O1…) et l'autre de la population P2 (M2N2O2…). La puissance de détection d'un QTL à la position x sur le génome est maximale lorsque l'origine, « 1 » ou « 2 », de la portion du génome reçue en x par un descendant est identifiée. En l'absence d'information sur des marqueurs, la probabilité π_x qu'un descendant reçoive la portion « 1 » du génome de son père en x est 1/2. Si l'on dispose d'information sur un seul marqueur M, situé à θ de la position x, π_x vaut $(1-\theta)$ si le descendant a reçu M1 et θ s'il a reçu M2. Si l'on dispose d'information sur deux marqueurs M et N, situés respectivement à θ_1 et θ_2 de part et d'autre de x, π_x peut être calculée selon le même principe en fonction de θ_1 et θ_2. Par exemple, en supposant qu'il n'y a pas d'interférence, π_x vaut $(1-\theta_1)(1-\theta_2)/((1-\theta_1)(1-\theta_2)+\theta_1\theta_2)$ si le descendant a reçu M1N1, $(1-\theta_1)\theta_2/((1-\theta_1)\theta_2+\theta_1(1-\theta_2))$ si le descendant a reçu M1N2, etc. La prise en compte de marqueurs flanquants permet donc à π_x , quand θ diminue, de tendre plus rapidement vers 1 qui représente l'information parfaite (le QTL est « sur » le marqueur ; figure 2).

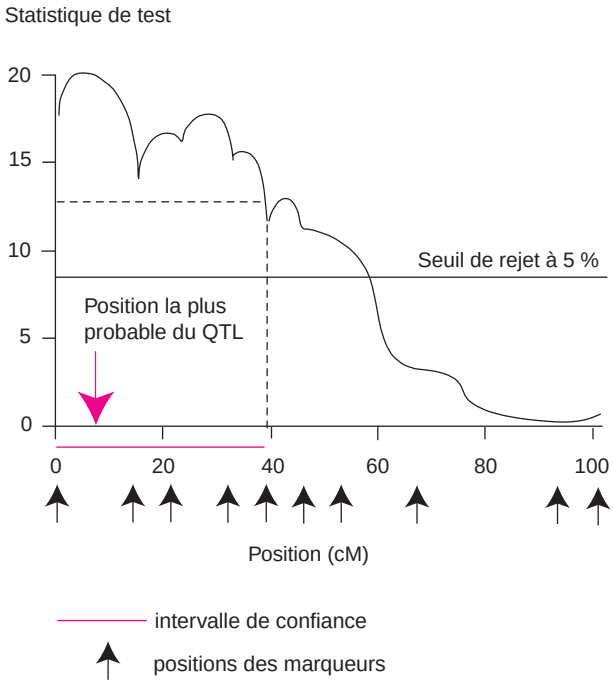
Figure 2. *Probabilité π_x qu'un descendant reçoive la portion « 1 » du génome de son père en x : sachant qu'il a reçu M1N1 (marqueurs flanquants à θ de x) ; sachant qu'il a reçu M1 (un seul marqueur à θ de x) ; sans information marqueur.*



Une fois les probabilités de transmission π_x calculées pour chaque descendant, la vraisemblance de l'échantillon V_x peut être estimée sous les deux hypothèses d'absence et de présence d'un QTL en x. La vraisemblance des observations considérée est la probabilité d'observer les phénotypes pour le caractère étudié (y) conditionnellement aux phénotypes aux marqueurs. Par exemple, dans le cas du CR, sous l'hypothèse qu'il existe un QTL à la position x, cette vraisemblance s'écrit comme le produit des vraisemblances des descendants k selon : $V_x=\Pi_k[\pi_x f_1(y_k)+(1-\pi_x)f_2(y_k)]$ où f_1 et f_2 sont les pénétrances des génotypes Q1Q1 et Q2Q1, en général modélisées par des lois normales de moyennes μ-a *et* μ+a et d'écart type commun σ. Les paramètres μ, a et σ sont estimés en maximisant V_x. L'hypothèse d'absence de QTL en x, qui correspond à l'hypothèse a=0, est testée contre l'hypothèse de présence d'un QTL en x avec un test du maximum de vraisemblance. Un profil de la statistique de test, le rapport de vraisemblance ou une approximation de celui-ci, est ensuite dressé pour le groupe de liaison (figure 3). La présence d'un QTL en x est affirmée si la statistique de test est maximum en ce point et supérieure au seuil de signification.

La démarche explicitée ici dans le cas simple d'un CR, peut être généralisée au cas des populations en ségrégation. Cependant, l'écriture de la vraisemblance de l'échantillon est alors plus compliquée. En effet, tous les marqueurs du groupe de liaison ne sont pas obligatoirement informatifs pour une famille donnée. Dans ce cas, le même principe est appliqué en prenant en compte les marqueurs informatifs adjacents les plus proches. Ceux-ci vont donc varier d'une famille à une autre. De plus, les haplotypes parentaux ne sont pas obligatoirement connus a priori et doivent être déduits, soit avec certitude à partir des données de marquage génétique sur les grands-parents si elles existent, soit en

Figure 3. *Exemple de profil de vraisemblance dans la cartographie d'intervalle : recherche de QTL d'épaisseur de lard dorsal sur le chromosome 4 du porc (d'après Andersson* et al *1994).*



probabilité à partir des données de marquage sur la descendance. Dans ce dernier cas, plusieurs haplotypes peuvent rester « possibles » pour un parent et cette incertitude doit être prise en compte.

Dans la pratique, différentes approximations numériques sont employées pour simplifier le calcul de la vraisemblance (Elsen *et al* 1999). L'idée principale est de généraliser l'approche par analyse de variance, utilisée pour un seul marqueur, en linéarisant la fonction de vraisemblance. La puissance de détection d'un QTL est relativement peu affectée par ces simplifications contrairement à la qualité des estimations des paramètres (μ, a, σ) qui tend à se dégrader. Par ailleurs, des approches non paramétriques, du type test de rang, ont également été proposées en remplacement de la démarche par maximum de vraisemblance (Coppieters *et al* 1998).

3 / Seuil de rejet et intervalle de confiance

Selon la théorie classique des tests d'hypothèse, si la valeur de la statistique de test calculée sur l'échantillon dépasse un certain seuil, l'hypothèse testée (absence de QTL dans le cas présent) est rejetée et par suite l'hypothèse générale est acceptée. Ce seuil de rejet dépend du risque de première espèce choisi par l'utilisateur du test : celui-ci « prend le risque » de rejeter à tort l'hypothèse testée avec une probabilité α. Dans le cas de la détection de QTL, α correspond donc au risque de trouver un QTL là où il n'y en a pas (« faux positif »). Par suite, le risque α choisi lors d'une analyse dépend de l'objectif fixé, soit être sûr des QTL trouvés (α faible), soit détecter tout QTL putatif (α élevé).

Lors de l'analyse des données des protocoles systématiques de détection de QTL, plusieurs centaines de tests sont réalisés au cours de la même expérience, chaque test correspondant à une position sur le génome (test de l'hypothèse « il n'y a pas de QTL à la position x »). Par exemple, si le seuil de rejet, choisi pour que la probabilité de ne pas avoir de faux positif soit de 95 % pour un test individuel, est employé dans 100 tests indépendants successifs, la probabilité de n'avoir aucun faux positif au niveau de l'expérience vaut $(0{,}95)^{100}$=0,6 %. Du fait de ces tests multiples, le risque de faux positif doit donc être raisonné, non pas pour chaque test individuel, mais au niveau global. Toutefois, même en tenant compte du nombre de tests pratiqués, le calcul du seuil de rejet n'est pas simple. En effet, la loi asymptotique du rapport de vraisemblance, dans le cas particulier de l'analyse de mélange de distributions, n'est pas une loi de χ^2 et les seuils de rejet pour un test ne sont par conséquent pas tabulés. Par ailleurs, les différents tests réalisés au niveau d'un groupe de liaison ne sont pas indépendants entre eux car les marqueurs sont liés. Enfin, dans la plupart des protocoles, plusieurs caractères sont étudiés, certains d'entre eux étant corrélés.

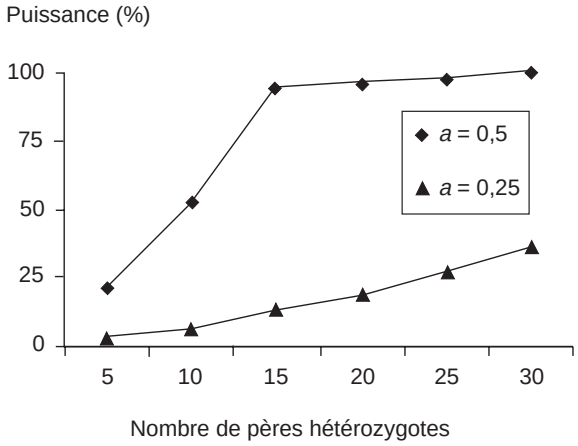
Ces différentes difficultés font qu'il n'y a pas de méthode exacte pour établir le seuil de rejet et être sûr du risque α réellement pris au cours d'une expérience. Dans le cas simple de croisements entre des lignées homozygotes, des formules algébriques, compliquées quoiqu'approximatives, ont été proposées pour calculer le seuil de rejet pour l'analyse d'un groupe de liaison (Rebaï *et al* 1994). En ce qui concerne les populations en ségrégation, la loi empirique suivie par la statistique de test sous l'hypothèse d'absence de QTL est en général établie par simulation (Churchill et Doerge 1994). Quelques milliers d'échantillons simulés sont alors nécessaires pour pouvoir estimer correctement le seuil de rejet à des niveaux faibles. Cette démarche implique des temps de calcul importants car elle doit être réitérée pour chaque groupe de liaison et chaque caractère. C'est la raison essentielle de l'emploi de simplifications numériques lors du calcul du rapport de vraisemblance. Par ailleurs, la difficulté du choix du niveau de rejet reste entière face au nombre important de tests réalisés : si les règles de calcul du seuil de rejet sont appliquées strictement, la puissance du dispositif devient rapidement très faible. Dans un article très philosophique, Lander et Kruglyak (1995) évoquent ce problème existentialiste et proposent une classification des QTL en QTL suggéré, significatif, hautement significatif et confirmé.

Lorsque l'existence d'un QTL en une position x est finalement retenue, il est indispensable de disposer d'un intervalle de confiance autour de x avant d'utiliser le QTL en sélection ou de le cartographier plus précisément. Comme pour le seuil de rejet, l'utilisation de techniques statistiques classiques pour construire cet intervalle de confiance est compliquée dans le cas des populations simples (Mangin *et al* 1994) et impossible dans le cas des populations en ségrégation. Les démarches mises en œuvre sont donc là aussi assez pragmatiques. La plus couramment employée consiste à prendre comme bornes de l'intervalle les positions entourant x auxquelles la probabilité d'existence du QTL est 10 fois moins grande qu'en x (Ott 1991 ; figure 3). L'utilisation de simulations a également été proposée pour une construction de l'intervalle de confiance par bootstrap (Visscher *et al* 1996).

4 / Protocoles

Comme nous l'avons vu, le principe de la détection de QTL est d'observer la descendance d'individus hétérozygotes aux marqueurs et aux QTL. La puissance de détection d'un protocole est donc directement liée au pourcentage de ces parents « informatifs » dans la population étudiée (figure 4). Dans le cas des animaux de laboratoire ou de certaines espèces végétales, l'existence de lignées consanguines permet de créer des parents hétérozygotes en de nombreux locus. Les protocoles consistent alors à produire des parents F1 puis des croisements de seconde génération, F2 ou CR, sur lesquels les caractères sont mesurés. Le CR sur le parent homozygote récessif pour le QTL est le dispositif le plus puissant mais aussi le plus risqué. En effet, si la dominance des allèles en un QTL est inversée par rapport à la dominance moyenne sur le génome, le CR est fait sur le parent homozygote dominant et il n'y a plus ségrégation des phénotypes chez les individus mesurés. Par ailleurs, si la dominance n'est pas complète, la différence de puissance entre F2 et CR est faible. Il est donc souvent préférable de produire des F2.

Figure 4. *Exemple d'évolution de la puissance d'un protocole en fonction du nombre de pères hétérozygotes au QTL (30 pères, 40 descendants par père, a égal à 0,5 ou 0,25 unité d'écart type phénotypique).*



En ce qui concerne les espèces d'élevage, les lignées homozygotes n'existant pas, la détection des QTL est toujours réalisée intra-famille. La puissance d'un protocole repose cependant sur les mêmes principes et peut être augmentée selon deux voies : maximiser le nombre de parents informatifs et maximiser la différence de performance entre groupes de descendants ayant reçu des allèles différents aux QTL.

Pour augmenter le nombre de parents informatifs, l'idée du croisement entre lignées homozygotes est reprise et appliquée à des populations extrêmes qui sont, soit des races très différentes quant à leurs performances (Bidanel et Milan 2000, cet ouvrage), soit des lignées divergentes sélectionnées (Pinard-van der Laan 2000, cet ouvrage). Il peut aussi s'agir d'un tri préalable dans une population en ségrégation de parents existants, déjà connus sur descendance et présentant des familles à grande variance pour les caractères étudiés.

Pour augmenter la différence entre groupes de descendants, il est possible de mettre en relation l'allèle marqueur reçu du parent hétérozygote avec, non plus la mesure du caractère, mais la valeur génétique des descendants pour le caractère. En effet, la variabilité autour d'une valeur génétique étant beaucoup plus faible que celle d'une mesure, il y a moins de recouvrement entre les distributions correspondant aux deux groupes de descendants ayant reçu l'un et l'autre allèle marqueur et, par suite, un même écart entre groupes est plus facilement visible. Ce protocole prévoit donc de marquer génétiquement des parents et leurs descendants et de mesurer les caractères sur les petits descendants pour prédire les valeurs génétiques des descendants, d'où son nom de protocole grand-père ou petits-descendants (Weller *et al* 1990). Il est parfaitement adapté pour les populations de bovins laitiers (Boichard *et al* 2000, cet ouvrage). Une autre possibilité pour augmenter la différence de performance entre groupes de descendants, à nombre de typages fixé, est de faire le marquage uniquement sur les descendants présentant les mesures les plus extrêmes (marquage sélectif) (Lander et Botstein 1989). Ceci implique toutefois de ne s'intéresser qu'à un seul caractère.

Conclusion

La détection des QTL grâce aux informations apportées par les marqueurs génétiques est depuis quelques années un thème de recherche en plein essor. Des méthodes statistiques spécifiques ont été développées pour analyser les premières données disponibles, et de nombreux QTL ont d'ores et déjà été mis en évidence dans différentes espèces. Toutefois, il reste de nombreux problèmes théoriques à résoudre pour répondre à toutes les questions posées. La maîtrise du taux d'erreur est notamment un défi urgent à relever. Par ailleurs, les modèles génétiques utilisés sont très simplistes par rapport à la complexité de la réalité biologique. Des résultats récents montrent par exemple l'existence d'empreintes génétiques, l'expression variable des QTL selon l'environnement, l'inversion des effets des allèles « sauvage » et « amélioré », etc. Il est donc certain que la notion même de QTL évolue et, avec elle, les perspectives de leur utilisation en élevage.

Références

Andersson L., Haley C.S., Ellegren H., *et al.*, 1994. Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science, 263, 1771-1774.

Bidanel J.P., Milan D., 2000. La recherche de QTL à l'aide de marqueurs : résultats chez le porc. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 223-228.

Boichard D., Grohs C., Bourgeois F., Cerqueira F., Faugeras R., Neau A., Milan D., Rupp R., Amigues Y., Boscher M.Y., Levéziel H., 2000. La recherche de QTL à l'aide de marqueurs : résultats

chez les bovins laitiers. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 217-222.

Churchill G.A., Doerge R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics, 138, 963-971.

Coppieters W., Kvasz A., Farnir F., Arranz J.J., Grisart B., Mackinnon M., Georges M., 1998. A rank based nonparametric method for mapping quantitative trait loci in outbred half sib pedigrees : application to milk production in a granddaughter design. Genetics, 149, 1547-1555.

Elsen J.M., 2000. Sélection et introgression assistées par marqueurs. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 233-237.

Elsen J.M., Mangin B., Goffinet B., Boichard D., Le Roy P., 1999. Alternative models for QTL detection in livestock. 1. General introduction. Genetics Selection Evolution, 31, 213-224.

Goffinet B., Beckmann J., Boichard D., *et al.*, 1994. Méthodes mathématiques pour l'étude des gènes contrôlant des caractères quantitatifs. Genetics Selection Evolution, 26, Suppl. 1, 9s-20s.

Lander E.S., Botstein D., 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121, 185-199.

Lander E., Kruglyak L., 1995. Genetic dissection of complex traits : guidelines for interpreting and reporting linkage results. Nature Genetics, 11, 241-247.

Mangin B., Goffinet B., Rebaï A., 1994. Constructing confidence intervals for QTL location. Genetics, 138, 1301-1308.

Ott J., 1991. Analysis of human genetic linkage. The John Hopkins University Press, Baltimore and London.

Pinard-van der Laan M.H., 2000. La recherche de QTL à l'aide de marqueurs : projets et résultats chez le mouton et la poule. INRA Productions Animales, numéro hors série « Génétique moléculaire : principes et application aux populations animales », 229-232.

Rebaï A. Goffinet B., Mangin B., 1994. Approximate thresholds of interval mapping tests for QTL detection. Genetics, 138, 235-240.

Soller M., Genizi A., 1978. The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. Biometrics, 34, 7-55.

Visscher P.M., Thompson R., Haley C.S., 1996. Confidence intervals in QTL mapping by bootstrapping. Genetics, 143, 1013-1020.

Weller J.L., Kashi Y., Soller M., 1990. Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. Journal of Dairy Science, 73, 2525-2537.